Human
Systems
Information
Analysis
Center

HSIAC-RA-2000-004

Review & Analysis

Auditory Workload Assessment, Volume I: Final Report

Prepared for: Human Research and Engineering Directorate
US Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5425

20010226 074

MICHAEL E. RENCH
Human Factors Analyst
Human Systems IAC

30 OCTOBER 2000

Human
Systems
Information
Analysis
Center

HSIAC-RA-2000-004

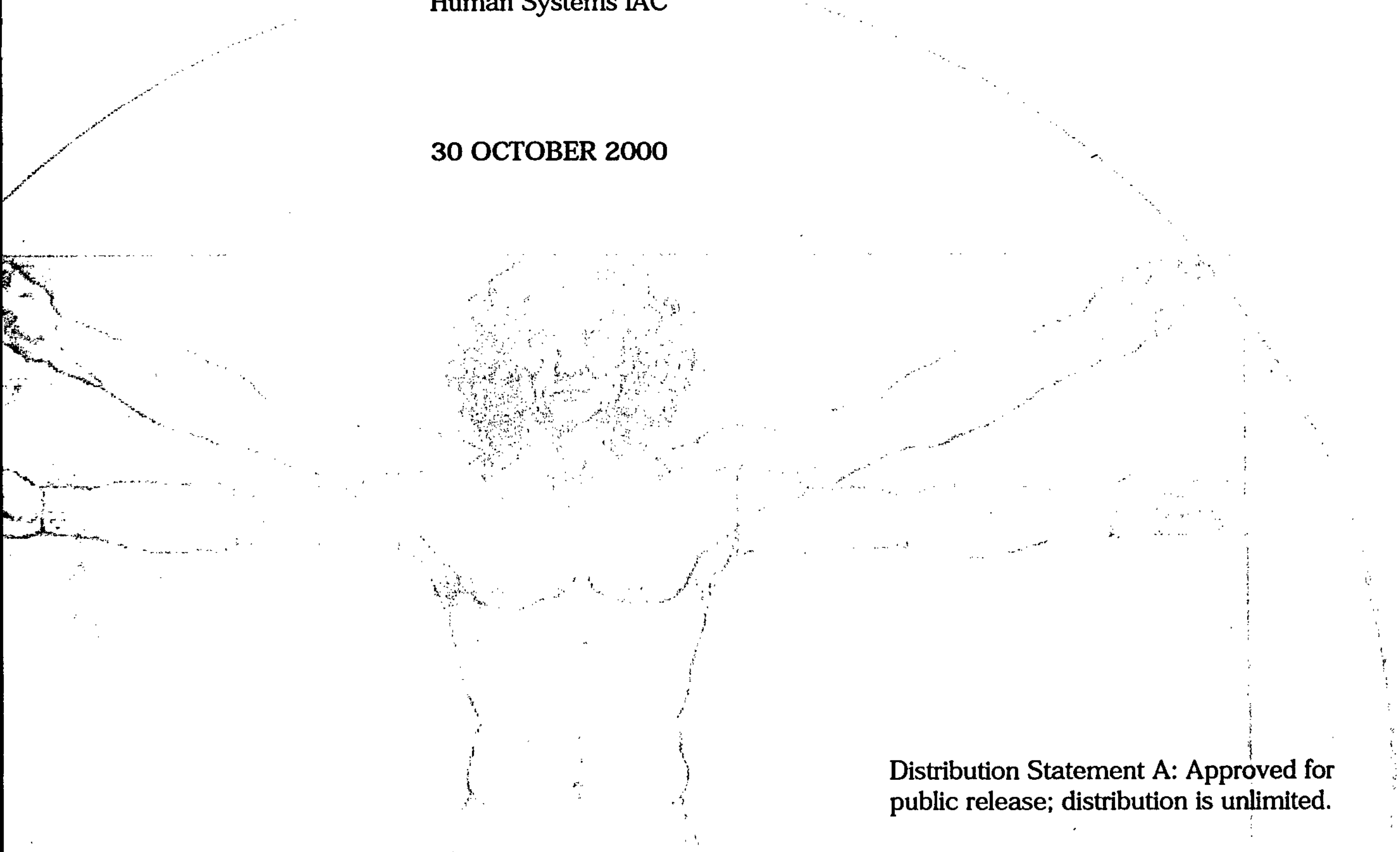
Review & Analysis

Auditory Workload Assessment, Volume I: Final Report

Prepared for: Human Research and Engineering Directorate
US Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5425

MICHAEL E. RENCH
Human Factors Analyst
Human Systems IAC

30 OCTOBER 2000



REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	10/30/00	Review & Analysis: Final Report 30 Oct 2000

4. TITLE AND SUBTITLE
Auditory Workload Assessment, Volume I: Final Report

5. FUNDING NUMBERS
SPO700-98-D-4001

6. AUTHOR(S)
Michael E. Rench

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

HSIAC Program Office
AFRL/HEC/CSERIAC Bldg. 196
2261 Monahan Way
WPAFB, OH 45433-7022
Commercial: (937) 255-4842 DSN: 785-4842

8. PERFORMING ORGANIZATION REPORT NUMBER
HSIAC-RA-2000-004

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Human Research and Engineering Directorate
US Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5425

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
This report contains three volumes. Volume I, this volume, is the final report on auditory workload assessment measurement. Volume II contains pertinent non-copyrighted citations extracted from government databases, and Volume III contains pertinent copyrighted citations extracted from commercial databases.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A: Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT ***(Maximum 200 Words)***

Human Systems IAC was asked to determine the existence of a measure designed specifically to assess auditory workload. Should no such measure be found, Human Systems IAC was then required to give recommendations for a research plan to develop such an auditory workload measure.

Human Systems IAC performed an in-depth literature search and consulted several subject matter experts to determine the existence of an auditory workload measure. Based on the available information, Human Systems IAC concluded that no auditory workload metric was available. As a result, recommendations were made for a new metric. Utilizing suggestions from related literature, it was recommended that an existing test, the NASA Task Load Index (TLX), be adapted to assess auditory workload. Suggestions were provided for this adaptation. Considerations for experimental design and selection of independent variables were also included in the methodology provided.

14. SUBJECT TERMS
Workload; Subjective Workload; Workload Assessment; Methodology; Audition; Auditory Workload

15. NUMBER OF PAGES
26

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

NOTICE

This report contains three volumes. Volume I, this volume, is the final report on auditory workload assessment measures. Volume II contains pertinent non-copyrighted citations extracted from government databases, and Volume III contains pertinent copyrighted citations extracted from commercial databases. A table of contents for the three-volume set may be found in Volume I.

EXECUTIVE SUMMARY

Human Systems IAC was asked to determine the existence of a measure designed specifically to assess auditory workload. Should no such measure be found, Human Systems IAC was then required to give recommendations for a research plan to develop such an auditory workload measure.

Human Systems IAC performed an in-depth literature search and consulted several subject matter experts to determine the existence of an auditory workload measure. Based on the available information, Human Systems IAC was unable to locate an auditory workload metric. As a result, recommendations were made for a developing a new metric by modifying an existing one. Utilizing suggestions from related literature, it was recommended that an existing test, the NASA Task Load Index (TLX), be adapted to assess auditory workload. Suggestions were provided for this adaptation. Considerations for experimental design and selection of independent variables were also included in the methodology provided.

VOLUME I
TABLE OF CONTENTS
FINAL REPORT

1. Introduction ... 1

1.1 Overview ... 1

1.2 Background ... 1

1.2.1 Workload ... 1
1.2.2 Workload Measures ... 2
1.2.3 Predictive Workload Models ... 3

1.3 Scope ... 3

2. Findings: Is there an Auditory Workload Metric? ... 3

2.1 Subject-Matter Expert Interviews ... 3

2.2 Literature Review ... 4

2.3 Summary of Findings ... 4

3. Recommendations for a Research Plan to Develop Auditory Workload Assessment Methodology ... 5

3.1 Plan Rationale ... 5

3.1.1 Using a New or Existing Measure ... 5
3.1.2 Selected Subjective Workload Measures ... 5

3.2 Proposed Research Plan ... 6

3.2.1 Adaptation of NASA-TLX to Assess Auditory Workload ... 6
3.2.2 Validation of the Adapted NASA-TLX ... 7

4. Conclusion ... 10

5. References ... 11

6. Appendix A: Literature Search Strategy ... 13

7. Appendix B: Search Terms ... 16

8. Appendix C: Example of Standard NASA-TLX Instructions ... 18

VOLUME II
TABLE OF CONTENTS
NON-COPYRIGHTED CITATIONS

1. Non-Copyrighted Literature Search Introduction 1

1.1 Introduction 1

1.2 Literature Search Strategy 1

1.2.1 Background 1
1.2.2 Search Terms 1
1.2.3 Key Authors 1

2. Database Searched 2

3. Ordering Documents 2

3.1 Defense Technical Information Center (DTIC) 2

3.2 National Technical Information Service (NTIS) 3

3.3 Human Systems Information Analysis Center (HSIAC) 3

4. Non-Copyrighted Literature Search 4

VOLUME III
TABLE OF CONTENTS
COPYRIGHTED CITATIONS

1. Copyrighted Literature Search Introduction 1

1.1 Introduction 1

1.2 Literature Search Strategy 1

1.2.1 Background 1
1.2.2 Search Terms 1
1.2.3 Key Authors 1

2. Databases Searched 2

3. Ordering Documents 2

3.1 Defense Technical Information Center (DTIC) 2

3.2 National Technical Information Service (NTIS) 3

3.3 Human Systems Information Analysis Center (HSIAC) 3

4. Copyrighted Literature Search 4

1. INTRODUCTION

1.1 OVERVIEW

This *Review & Analysis* (*R&A*) begins with a background discussion of the problem addressed, relevant workload topics, and workload measures. It then covers findings from subject matter expert (SME) interviews and an in-depth literature review. The recommendations for the selection, development, and testing of a research methodology to assess auditory workload are then presented. The document closes with a brief conclusion.

1.2 BACKGROUND

Auditory research has shown that the use of different types of auditory displays (e.g., monaural vs. 3-D audio) result in differences in human performance. Several Army Research Laboratory (ARL) studies have shown that using 3-D audio displays versus monaural displays allow the operator to process a significantly greater number of target messages with a significantly shorter response time. However, traditional subjective measures of mental workload (e.g., SWAT, NASA-TLX) have revealed no corresponding difference in experienced operator workload. The question is raised whether the metrics used are sensitive to the specific demands of auditory processing (E. Haas, personal communication, 15 August, 2000).

Although several models exist to predict the sensory components of workload (Sarno & Wickens, 1995), ARL is not aware of a workload measure used to assess the workload demands directly associated with auditory processing. The development of such a measure could be used to establish a relationship between audio display design and soldier workload. This workload assessment could then be applied to reduce workload and, consequently, enhance soldier performance.

1.2.1 Workload

In an excellent tutorial, O'Donnell and Eggemeier (1986) define workload as the term used to describe the portion of an operator's limited mental capacity required to perform a particular task, given that increases in task difficulty lead to increases in resource expenditure. No matter what definition is used, the goal of a system designer is usually to achieve "optimal" workload. Optimal workload is defined by Hart (1991) as "a situation in which the operator feels comfortable, can manage task demands intelligently, and maintain good performance" (p. 3). Optimal workload enables an operator to perform at his/her full potential (Hart, 1991). The ability of a soldier to operate at an optimal level should enhance his/her performance. Consequently, better performance should improve effectiveness.

There are as many theories to explain workload as there are definitions. One theory (Kahneman, 1973) suggests that workload is the drain on a system's single store of "processing resources." These resources come from a single undifferentiated "pool" of energizing forces needed to complete a task. Another theory (Wickens, 1991) takes a somewhat different approach in supporting the notion that these resources exist, but that they are differentiated among multiple "stores" of resources. The strongest empirical support is for a multiple capacity model (Shingledecker, Crabtree, & Acton, 1982). One of the most accepted multiple capacity theories is Wicken's multiple resource theory (MRT) (1991). MRT suggests that humans have a limited capacity for processing information. Therefore, if an operator must perform multiple tasks at the same time, performance on one or all of the tasks may suffer. This is because each task has less resources devoted to it than if it were performed separately (Mitchell, 2000).

1.2.2 Workload Measures

Measures of mental workload can be divided into three broad areas: physiological, performance/behavioral, and subjective measures (Mitchell, 2000; Whitaker, Hahus, & Birkmire-Peters, 1997; Shingledecker, Crabtree, & Acton, 1982). Wilson and Eggemeier (1994) and O'Donnell and Eggemeier (1986) provide excellent guidance in the selection and use of the various workload measures.

1.2.2.1 Physiological Measures

The human body responds physically and cognitively to the demands imposed by tasks. Some measures of physiology vary directly with cognitive demands. These potential metrics (e.g., eye blink rate, heart rate, pupil diameter, P300 amplitude and latency, and galvanic skin response) can be tested. For instance, heart rate is expected to increase as workload increases. Therefore, if an operator's heart rate increases while performing a task, it is likely that the task is increasing his/her level of workload. However, measures often do not agree with one another (i.e., a task demand may be reflected in P300, but not in heart rate) and agreements between measures do not occur consistently in the literature (Whitaker, Hahus, & Birkmire-Peters, 1997). Physiological measures also do not correlate well with performance, however, they do help identify areas of high workload that may impact performance. These areas can be addressed by designers before a system is fielded (Mitchell, 2000). (See Kramer, 1991, for an extensive review of physiological measures of workload.)

1.2.2.2 Performance/Behavioral Measures

Performance or behavioral measures of workload are based on the assumption that as an operator's workload increases, his/her performance on a task decreases. These metrics are often employed in field settings. Performance measures are typically divided into two types of metrics, primary and secondary task measures. Primary task measures examine the operator's ability to perform a required task in a given system (e.g., fly a straight line in a simulator). Secondary task measures augment the primary task methodology by asking the participant to perform a concurrent task. This "secondary task" is designed to utilize the operator's reserve processing capability (Mitchell, 2000). Although the primary focus of performance measures are changes in cognitive workload, Whitaker, Hahus, and Birkmire-Peters (1997) state that, "Performance is not a sensitive indicator of the changes in cognitive workload," (p. 5) unless the task performance is sensitive to changes in workload. This may be due to the tentative relationship between performance and mental workload. For example, a task requiring low mental resources can be performed well by an operator. If the task demands are increased, the person may have the same performance, but experience a higher level of workload.

1.2.2.3 Subjective Measures

Subjective measures of workload are instruments designed to measure an operator's personal evaluation of the difficulty of a task. These measures have achieved the greatest success of all of the empirical techniques of assessing workload by simply asking the operator to asses his/her own mental workload (Moray, 1988, cited in Mitchell, 2000). These tests are also easier to administer than the other metrics since they can be given after a task using a pencil and paper. Physiological and performance tests typically require more complex apparatus and can have greater interfere with

the primary focus of a study. When compared to physiological tests and task performance, subjective measures are generally sensitive, reliable, and have high face validity (Whitaker, Hahus, & Birkmire-Peters, 1997). Face validity means that a measure looks as though it measures what it is intended to measure (Sanders & McCormick, 1993).

1.2.3 Predictive Workload Models

In addition to three areas of mental workload measurement, there are also several models that are used to help *predict* the levels of workload that an operator will experience when using a given interface. These models are often used early in the development process of various interfaces (e.g., cockpits, command and control vehicles, etc.) by predicting the various kinds of workload that an operator might experience (e.g., visual, auditory, cognitive, temporal, etc.). Although not applicable to the development of a measure of auditory workload, there was some question regarding the applicability of these models in developing a measure. However through our research (Sarno & Wickens, 1992; Aldrich, Szabo & Bierbaum, 1988; North & Riley, 1988; Parks & Boucek, 1988) it became clear that these models are more like calculated estimates of the eventual operator workload and do not actually measure it. Cohen, Wherry, and Glenn (1993) state that, "these estimates may not be valid indications of the real effort levels that will be required of operators when the actual system has been developed."

1.3 SCOPE

Human Systems IAC is tasked with performing a literature search and contacting subject-matter experts (SMEs) to determine whether or not metrics or measurements of auditory workload demands exist. If they do exist, Human Systems IAC will then identify these measures and assess their relevance to the measurement of soldier performance. If no such measures exist, Human Systems IAC will make recommendations for the development of a research plan whose implementation would result in such a methodology.

2. FINDINGS: IS THERE AN AUDITORY WORKLOAD METRIC?

2.1 SUBJECT-MATTER EXPERT INTERVIEWS

Human Systems IAC interviewed six subject-matter experts (SME) who have conducted or are currently conducting work in the fields of workload or audition. The SMEs were selected based on the literature review, suggestions made by the customer, and recommendations from other experts. The goal of the interviews was to acquire the most current information regarding the possible existence of a measure of auditory workload and provide any information that may not have yet been published. Interviews were conducted over the telephone and averaged approximately 15 minutes each.

All of the experts were asked the same basic question, "Do you know of a method or scale for measuring workload demand associated with auditory processing?" Discussions of varying length ensued as a result of this question. While the information gathered was interesting and informative to the author of this *R&A*, in the end, the answer from every expert was, "No" (see Table 1). None of the experts listed in Table 1 had worked on or knew of a metric for auditory workload. In fact, the first response made by every person interviewed was, "Why would you need one?" This question was asked in reference to the generally accepted nature of workload as having a global

impact on the operator. The typical use for a workload measure is to determine overall capacity rather than assess one sensory area.

Table 1. Contributing SMEs

Expert	Area of Expertise	Know of AW Measure?
Robert Bolia	Auditory display RDT&E	No
F. Thomas Eggemeier, Ph.D.	Mental workload	No
Mark Ericson	Perception/Communication engineering	No
William F. Moroney, Ph.D.	Workload/Human factors	No
David Payne, Ph.D.	Mental workload/Cognitive psychology/ Performance assessment measures	No
Michael Vidulich, Ph.D.	Mental workload/Cognitive psychology/ Performance assessment measures	No

2.2 LITERATURE REVIEW

Human Systems IAC also conducted an in-depth literature review (see Appendix A: Literature Search Strategy). The results of that search provided the bulk of the background resources for this document. All relevant sources resulting from that search can be found in Volumes II and III of this *Review & Analysis* (*R&A*). The MATRIS office of the Defense Technical Information Center (DTIC) also conducted a search to augment the internal Human Systems IAC findings. The results both searches can be found in Volumes II and III.

2.3 SUMMARY OF FINDINGS

Based on the survey results of several SMEs and an extensive literature search, Human Systems IAC was unable to find a specific metric for auditory workload processing. While Human Systems IAC found two general areas of research that were related, general workload measures and predictive models, neither included a specific measure. Although several workload metrics that include a mental and/or physical component were identified (e.g., NASA-TLX, SWAT, Cooper-Harper), none of them included a sensory-specific aspect (see Section 1.2.2.3). Another area that had potential for including an auditory workload measure was workload modeling. Several predictive models of workload include an auditory component (e.g., TLAP, VACP, W/INDEX), however these tests are geared toward the prediction of workload based on the recommendations of what experts expect to experience in a situation, and not on direct measurement (See Section 1.2.3).

3. RECOMMENDATIONS FOR A RESEARCH PLAN TO DEVELOP AUDITORY WORKLOAD ASSESSMENT METHODOLOGY

3.1 PLAN RATIONALE

The next two sections outline the rationale behind the proposed research plan.

3.1.1 Using a New or Existing Measure

Since Human Systems IAC was unable to locate an existing auditory workload measure, the IAC must make recommendations for a research plan to develop an auditory workload measure. As a first step in preparation for making such recommendations, it must be determined whether it is more logical to develop a mew measure or modify an existing one. The first option, designing an entirely new workload measure, could be a challenging undertaking. The designers would have to develop or adapt new ways to measure auditory workload and methods to scale the information gathered so that it is useful. The new metric would then have to be validated and proven reliable, sensitive, selective, and acceptable to the user community (Shingledecker, Crabtree, & Acton, 1982; Sanders & McCormick, 1993).

The second option, to modify an existing workload measure to assess auditory workload, may be easier than developing a new metric. Changing an existing measure would only require the designers to change the established instructions to reflect the new focus on auditory workload. Many of the inherent problems associated with developing a new measure would be avoided. For instance, subjective measures of workload are susceptible to contamination by experimenter and participant expectations. Any new measure would have to be tested for this effect and adjusted accordingly (Whitaker, Hahus, & Birkmire-Peters, 1997). While verification and validation of this modified metric would be necessary, it should retain most of its validity and reliability. Some preliminary pilot testing would be required to establish the modified measure's sensitivity and selectivity (F. T. Eggemeier, personal communication, October 4, 2000).

Based on the information available, Human Systems IAC recommends that the customer modify an existing measure.

3.1.2 Selected Subjective Workload Measures

Given the recommendation to modify an existing measure, the next step is to select the best measure to modify. While there are three broad areas of workload measures to choose from (see Section 1.2.2), at the request of our customer this *R&A* will focus on subjective measures. Although there are many subjective measures of workload available, there are only three methods recommended by Whitaker, Hahus, and Birkmire-Peters (1997) as having the most theoretical support and the highest ratings in eight categories for successful mental workload metrics. These categories are summarized in the five usefulness criteria (Sanders & McCormick, 1993) listed in Section 3.2. These three measures are the Subjective Workload Assessment Test (SWAT) (Reid, et al. 1981), the Cooper-Harper Scale (Cooper & Harper, 1969), and NASA Task Load Index (NASA-TLX) (Hart & Staveland, 1988). All three assessment tools are multidimensional, that is they address different components of workload, and are valid and sensitive enough to be utilized in the proposed research plan. They also have been modified to prevent contamination by experimenter and participant expectations (Whitaker, Hahus, & Birkmire-Peters, 1997).

SWAT, developed by Reid, et al. (1981), is a subjective measure of mental workload that divides the operator's resources into three intuitively derived dimensions; time load, mental effort load, and psychological stress. Participants rate their workload on a three-point scale across each of

the three dimensions. The result is a single score of operator workload. Although the three dimensions have not been empirically validated and have been shown to be somewhat interdependent (Boyd, 1983), SWAT has been found to be a valid, reliable, and sensitive measure of mental workload (Whitaker, Hahus, & Birkmire-Peters, 1997).

The Cooper-Harper Scale is another subjective workload metric that might be modified to measure auditory workload. It was originally designed to assess workload experienced by pilots in the cockpit relative to the aircraft handling qualities. The Scale applies a decision tree and a 10-point rating scale (scores vary from 1: very easy, through 5: moderately difficult, through 10: impossible) to develop a workload score. With minimal rewording the metric is a sensitive measure for many motor and psychomotor tasks as well as perceptual, cognitive, and communications tasks (Sanders & McCormick, 1993; Cooper & Harper, Jr., 1969).

The Task Load Index developed by NASA Ames Research Center is the third metric that could be used in assessing auditory workload. NASA-TLX provides an overall workload score based on operator ratings of six subscales: mental demands, physical demands, temporal demands, own performance, effort, and frustration. The overall score is often based on weighted averages, however there is some question regarding the value of this extra step (Moroney, Biers, & Eggemeier, 1995). NASA-TLX produces consistent, reliable subjective workload rating scores (Sanders & McCormick, 1993). NASA-TLX has been shown to be a valid, reliable, and sensitive measure of cognitive workload (Whitaker, Hahus, & Birkmire-Peters, 1997).

While all three measures have been shown to be capable measures and could be adapted to meet the needs of the customer, the customer has determined that NASA-TLX is the best choice for them to modify into an auditory workload metric. This decision is based on the proven history of NASA-TLX as a consistent and reliable subjective workload rating as well as its immediate availability to the customer.

3.2 PROPOSED RESEARCH PLAN

The proposed research plan for adopting NASA-TLX into an auditory workload measure will consist of two components. The first will be to recommend adapting the NASA-TLX from a globally oriented workload measure to one designed to specifically assess auditory workload. The second component will be to validate the adapted NASA-TLX as a viable, useful tool.

3.2.1 Adaptation of NASA-TLX to Assess Auditory Workload

The first step in adapting NASA-TLX to an auditory workload metric should be to change the instructions for NASA-TLX to focus on the auditory component of workload. This is necessary because NASA-TLX was designed to measure the global mental workload experienced by an operator, not just one sensory component. Appendix C gives an example of the standard instructions used by a global NASA-TLX survey. The instructions should be altered to reflect the new focus desired. For example, instead of referring to a general "task" or an individual's global "experience," the phrases should be changed to "auditory component of the task" or just "auditory task," and the term "experience" should be changed to "auditory experience," respectively.

These are just a few examples of the changes necessary. The experimenter should make the final adjustments and ensure that the new instructions make sense to the operator and focus his/her attention appropriately. It may even be worth the additional step of instructing the operator to pay special attention to his/her auditory experiences prior to the start of the experimental conditions. While this is different from the standard NASA-TLX methodology, it would provide for the unique circumstances and help focus the operator's observations accordingly.

3.2.2 Validation of the Adapted NASA-TLX

After NASA-TLX has been adapted to assess just auditory workload, a methodology for assessing its validity must be developed. Validity is the extent to which a metric measures what it was supposed to measure (Sanders & McCormick, 1993). In addition to verifying the validity of the adapted NASA-TLX, a test should also be designed to ensure the usefulness of the metric as described by Sanders and McCormick.

For the metric to be be useful, it should tell the experimenter something he/she did not already know. Shingledecker, Crabtree, and Acton (1982) and Sanders and McCormick (1993) describe a useful mental workload metric as having five basic criteria. These criteria will be used as the organizational framework for the recommended experimental testing of the validity of the adapted NASA-TLX:

1. Sensitivity: the measure should distinguish task situations that intuitively appear to require different levels of mental workload.
2. Selectivity: the measure should not be impacted by things not generally considered to be part of mental workload, such as physical or emotional stress.
3. Interference: the measure should not interfere with or contaminate the primary task that is being assessed.
4. Reliability: the measure should be reliable and repeatable over time (test-retest reliability).
5. Acceptability: the measuring technique should be acceptable to the person being measured.

In addition to the usefulness criteria, there are two additional considerations that should be addressed when selecting independent variables for experiments to validate an auditory workload assessment tool. These areas are overall system-oriented variables and specific auditory-oriented variables.

Meister (1999) discusses three types of system-oriented variables that can be used to describe a system's characteristics: general system variables, system structural variables, and general behavioral variables. These variables should also be considered when measuring a system. General system variables include factors such as requirements, functions, mission, and goals. System structural variables include characteristics like system size, number of subsystems, system complexity, transparency, autonomy, and dependency. The general behavioral variables focus on factors such as tasks performed by personnel, personnel experience/skill requirements, physical environment, and factors leading to performance degradation.

When selecting specific auditory-oriented variables, three factors should be addressed: transmission factors, linguistic factors, and individual factors (Peters, 1991). Transmission factors include the intelligibility and structure of the message being received. Intelligibility is the percent of correctly identified messages (e.g., signal-to-noise ratio) and structure is a combination of the number of exchanges and paths of communication within a given level of intelligibility (e.g., command, interrogative, discussion). Linguistic factors include the criticality, expectancy, and complexity of a message with respect to the operator. Criticality can be defined as the need for the information. Expectancy describes how prepared the operator is for the information. Complexity is the degree of interaction of various linguistic rules in a message. Individual factors of auditory-oriented variables are made up of the resources that the operator brings to the situation. These include training, experience, and personal ability (Peters, 1991).

In addition to the overall research recommendations for experimental organization and independent variable selection provided above, Human Systems IAC has also included some suggestions for initial testing of hypotheses. Below are some example tests that could be used to validate the usefulness of the modified NASA-TLX. While they focus primarily on individual

factors and general behavioral variables, all appropriate aspects of the system variables and auditory factors should be addressed in the final research program.

3.2.2.1 Testing of Sensitivity

The purpose of the first group of studies would be to determine if the new instructions of the adapted NASA-TLX are effective. They will also determine the gross sensitivity of the metric. The method of measuring sensitivity should focus on establishing clear levels of workload based on some empirical reasoning. The adapted NASA-TLX can then be tested against those levels. For example, utilize workload tasks with two or more levels of task difficulty (Keppel, 1982) that have been shown to have different, distinct levels in previous studies. Some areas that should be investigated would be number of signals (see example study below), sound intensity (decibel level), and interaction with background noise (Peters, 1991).

Example Sensitivity Study:

Hypothesis: Given that the adapted NASA-TLX is a sensitive metric, it will accurately detect changes in tests with clearly distinct levels of auditory workload and provide a numerical score for the levels.
N: 15+ (Keppel, 1982)
IV: Levels of auditory task difficulty
DV: Scores on the adapted NASA-TLX
Apparatus: Adapted NASA-TLX; tasks with levels of auditory workload that can be varied while keeping all other aspects of task the same. We will use a flight simulator example.
Procedure: After the usual experiment beginning (consent forms, etc.), give the participant any pre-instructions, if applicable (see Section 3.2.1 for a discussion on prior instructions), administer a workload-oriented flight simulator task. Keep all of the stimuli (visual, haptic, vestibular, etc.) consistent and vary only the auditory input. The auditory input should vary in a clear and intuitively obvious way. For instance, low, medium, and high workload with the operator handling one, five, and ten signals per minute respectively. At the end of the task, administer the adapted NASA-TLX and record the results. Perform necessary statistical analyses.
Statistical analysis: One-way, within subjects (repeated measures) Analysis of Variance (ANOVA) or parametric ANOVA, depending on the data.
Results: Theoretically, the results will correlate directly with the expected levels of workload.

3.2.2.2 Testing of Selectivity

The purpose of this group of studies is to further assess the validity of the adapted NASA-TLX by varying the non-auditory stimuli. If the visual, physical, etc. stimuli are changed while the auditory stimulus remains the same, the adapted NASA-TLX results should reflect no change. These tests also check the selectivity of the new metric. Special attention should be paid to the channel chosen as an independent variable. Multiple resource theory (see Section 1.2.1) suggests that similar channels will have more impact on mental resources than dissimilar channels (Wickens, 1984). Some stimulus areas that should be addressed are visual, physical, tactile, and vestibular (Mitchell, 2000).

Example Selectivity Study:

Hypothesis: Given that the adapted NASA-TLX is a selective measure, any increase/decrease in sensory input other than auditory will not impact the auditory test results.
N: 15+ (Keppel, 1982)
IV: Intensity of external stimuli
DV: Scores on the adapted NASA-TLX
Apparatus: Adapted NASA-TLX; some workload task that can vary levels of external stimuli (e.g., visual complexity), while maintaining a consistent level of auditory workload. We will continue to use the flight simulator example.
Procedure: After the usual experiment beginning (consent forms, etc.), give the participant any pre-instructions, if applicable (see Section 3.2.1 for a discussion on prior instructions). Then administer a workload-oriented flight simulator task. Keep auditory workload levels as consistent as possible (e.g., five signals per minute for all conditions) and vary one or more aspects of the remaining stimuli (visual, haptic, vestibular, etc.). Administer the adapted NASA-TLX after the task and record the results. Perform necessary statistical analyses.
Statistical analysis: One-way within subjects (repeated measures) Analysis of Variance (ANOVA) or parametric ANOVA, depending on the data.
Results: Theoretically, the results of the auditory workload scores should remain the same, regardless of the varying sensory input.

3.2.2.3 Testing of Reliability

The purpose of the remaining studies would be to further refine the methodology used in applying and scoring the adapted NASA-TLX. Continued tests would also be effective at determining the reliability of the new measure by re-testing the methodology in varying situations. The experimental components (e.g., number of subjects, apparatus, procedure) should mirror the suggested experimental designs presented in the previous two sections. It should be noted that the example studies focus on individual factors with general behavioral variables. Continued research should incorporate all aspects of the system (i.e., general system variables, system structural variables) and auditory factors (i.e., transmission, linguistic) contributing to workload as needed.

3.2.2.4 Addressing Interference and Acceptability

The described research plan and associated preliminary studies have addressed three of the five criteria for a useful mental workload metric. The final two, interference and lack of acceptability, are obviated by the nature of the NASA-TLX metric. Interference should not be a factor as NASA-TLX or the adapted NASA-TLX is employed after the task is completed. It is unlikely for the metric to interfere with the task. Acceptability should not be an issue since the test is relatively benign and simple to take/administer. The standard form of NASA-TLX has been in use for years.

4. CONCLUSION

This *Review & Analysis* described the research conducted, the findings, and our recommendations for the development of a modified measure for auditory workload. Human Systems IAC began this effort by questioning subject-matter experts and conducting an in-depth literature review. Based on this review, it was established that, given the information available, an auditory workload metric does not exist. As a result, Human Systems IAC then provided recommendations for the selection, development, and testing of a modified metric.

If this new method of adapting NASA-TLX to measure the auditory component of workload is shown to be effective, the method itself could lead to an entirely new battery of sensory-oriented tests of workload. Metrics for all types of sensory input could be developed quickly and at a relatively low cost. These tests could then be employed to completely validate predictive models and existing areas that require measurement of sensory workload.

5. REFERENCES

Aldrich, T., Szabo, S., & Bierbaum, C. R. (1988). The development and application of models to predict operator workload during system design. In G. McMillan, et. al (Eds.), *Applications of human performance models to system design* (pp. 65-80). New York: Plenum Press.

Boyd, S. (1983). Assessing the validity of SWAT as a workload measurement instrument. In *Proceedings of the Human Factors Society 27th Annual Meeting* (pp. 124-128). Santa Monica, CA: Human Factors Society.

Cohen, D., Wherry, R. J. Jr., & Glenn, F. (1993). *A critical analysis of workload prediction generated by multiple resource theory during early crewstation design* (Report No. NAWCADWAR-TN-93043-60; XB-NAWCADWAR-60). Warminster, PA: Naval Air Warfare Center, Aircraft Division. (DTIC No. ADA274804)

Cooper, G. & Harper, R. (1969). *The use of pilot ratings in the evaluation of aircraft handling qualities* (NASA Ames Tech Report No. NASA TN-D-5153). Moffett Field, CA: NASA Ames Research Center.

Hart, S. G. (1991). *Pilots' workload coping strategies*. Paper presented at the AIAA/NASA/FAA/HFS Confrence on challenges in aviation human factors, Tyson's Corner, VA, January.

Hart, S. & Staveland, L. (1988). Development of NASA-TLX: Results of empirical and theoretical research. In P Hancock and N. Meshkati (Eds.), *Human mental workload.* Amsterdam: North-Holland.

Kahneman, D. (1973). *Attention and effort.* Englewood Cliffs, NJ: Prentice-Hall.

Keppel, G. (1982). *Design and analysis: A researcher's handbook.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Kramer, A. F. (1991). Physiological metrics of mental workload: A review of recent progress. In D. L. Damos (Ed.), *Multiple task performance* (pp. 279-328). Washington, DC: Taylor & Francis.

Mitchell, D. K. (2000). *Mental workload and ARL workload modeling tools* (Report No. ARL-TN-161; XA-ARL/APG). Aberdeen Proving Ground, MD: Army Research Lab. (DTIC No. ADA377300)

Moroney, W. F., Biers, D. W., & Eggemeier, F. T. (1995). Some measurement and methodological considerations in the application of subjective workload measurement techniques. *International Journal of Aviation Psychology, 5,* 87-106.

NASA Task Load Index (TLX) computerized version 1.0 (1986). Moffett Field, CA: NASA Ames Research Center.

North, R. & Riley, V. (1988). W/INDEX: A predictive model of operator workload. In G. McMillan, et. al (Eds.), *Applications of human performance models to system design* (pp. 81-89). New York: Plenum Press.

O'Donnell, R. & Eggemeier, F. T. (1986). Workload assessment methodology. In K. Boff, L. Kaufman, & J. Thomas (Eds.), *Handbook of perception and human performance, Volume II: Cognitive processes and Performance* (42-1 – 42-49).

Parks, D. L. & Boucek, G. P., Jr. (1988). Workload prediction, diagnosis, and continuing challenges. In G. McMillan, et. al (Eds.), *Applications of human performance models to system design* (pp. 65-80). New York: Plenum Press.

Peters. L. J. (1991). Auditory performance. A model to predict task performance as a function of auditory workload. Overview. *In Proceedings of the Human Factors Society 35th Annual Meeting*, San Francisco, CA (pp. 609-613). Santa Monica, CA: Human Factors Society.

Reid, G. B. Shingledecker, C. A., Nygren, T. E., & Eggemeier, F. T. (1981). Development of multidimensional subjective measures of workload. In *Proceedings of the 1982 IEFF International Conference of Cybernetics and Society* (403-406).

Sanders, M. S. & McCormick, E. J. (1993). *Human factors in engineering and design.* New York, NY: McGraw-Hill, Inc.

Sarno, K. J. & Wickens, C. D. (1995). Role of multiple resources in predicting time-sharing efficiency: Evaluation of thee workload models in a multiple-task setting. *International Journal of Aviation Psychology, 5*(1), 107-130.

Shingledecker, C. A., Crabtree, M. S., & Acton, W. H. (1982). Standardized test for the evaluation and classification of workload metrics. *In Proceedings of the Human Factors Society 26th Annual Meeting*. Santa Monica, CA: Human Factors Society.

Whitaker, L. A., Hahus, J., & Birkmire-Peters, D. (1997). Selection of a workload metric for evaluation of telemedicine applications: Literature review and methodological development.

Wickens, C. D. (1984). *Engineering psychology and human performance*. Columbus, OH: Merrill.

Wickens, C. D. (1991). Processing resources and attention. In D. L. Damos (Ed.), *Multiple task performance* (pp.3-34). Washington, DC: Taylor & Francis.

Wilson, G. F. & Eggemeier, F. T. (1994). Mental workload assessment. *CSERIAC Gateway, 5*(2), 1-4.

6. APPENDIX A: LITERATURE SEARCH STRATEGY

Auditory Workload *Review & Analysis*

Literature Search Strategy

For: Army Research Laboratory
Aberdeen Proving Ground, MD

Background:

Human Systems IAC has been asked to prepare a *Review & Analysis* on workload measures or scales that can be used to assess the workload demands associated with auditory processing. This effort stems from ARL studies showing that 3-D audio displays allow operators to process a significantly greater number of target messages in a significantly shorter time than with traditional monaural displays. However, the traditional measures of workload (NASA TLX [Task Load Index] and SWAT [Subjective Workload Assessment Technique]) do not indicate any difference in the level of workload between the two displays. This information suggests that these results are due to a lack of sensitivity in the scales used. Therefore, a suitable and valid measure of auditory workload must be identified and employed. The identification of this measure is the primary goal of this literature search.

It should be noted that ARL is not aware of any measurement that can assess workload demands associated with auditory processing. Therefore this search may be an effort to determine what is *not* out there. As a result, special attention needs to be made regarding the methodology of the search to prevent Human Systems IAC from missing any major sources of information. However, should no auditory workload scale be available, Human Systems IAC will be responsible for providing recommendations for the development of a research plan that would result in such a methodology that is capable of detecting auditory processing demands.

Search Terms:

See Appendix B.

Key Authors:

David G. Payne
Christopher D. Wickens
F. Thomas Eggemeier
Mark Erickson
Richard L. McKinley
Leslie J. Peters (to see her references)

Databases Used:

Aerospace Database
Cambridge Scientific Abstracts
Defense Technical Information Center (DTIC)
ISI Science Citation Index
NASA Recon
NTIS
PsychINFO
Web of Science

Example Articles:

Backs, R. W. & Walrath, L. C. (1991). Heart rate variability and auditory workload during noise stress - Speaker sex and bandpass effects on speech intelligibility (Report No. A92-44901 19-53). In *International Symposium on Aviation Psychology, 6th, 2*, (pp. 740-745). Columbus, OH, Apr. 29-May 2. Columbus, OH: Ohio State University.

Birkmire, D. P. (1991). Auditory Performance: A Model to predict performance as a function of auditory workload: Linguistic factors. In *Visions. Proceedings of the human Factors Society 35th Annual Meeting, 1,* (pp. 618-621). San Francisco, CA, September 2-6. Santa Monica, CA: The Human Factors Society.

Fowler, B. (1994). P300 as a measure of workload during a simulated aircraft landing task. *Human Factors, 36*(4), 670-683.

Payne. D. G., Peters L. J., Birkmire D. P., Bonbo, M. A., Anastasi, J. S., & Wenger, J. J. (1994). Effects of speech intelligibility level on concurrent visual task-performance. *Human Factors, 36*(3), 441-475.

Peters, L. J. (1991). Auditory performance: A Model to predict task performance as a function of auditory workload: Overview. In *Visions. Proceedings of the human Factors Society 35th Annual Meeting, 1,*(pp. 609-613). San Francisco, CA, September 2-6. Santa Monica, CA: The Human Factors Society.

Sarno, K. J. & Wickens, C. D. (1995). Role of multiple resources in predicting time-sharing efficiency: Evaluation of three workload models in a multiple-task setting. *International Journal of Aviation Psychology, 5*(1), 107-130.

7. APPENDIX B: SEARCH TERMS

	ARL AUDITORY WORKLOAD *R&A* SEARCH STRATEGY						
	Main Ideas		Secondary Terms		R&A/Search Focus		Additional Items of Interest
1	Auditory Audition Hearing Audit*	A N D	Processing Work Effort Demands Detection Coding Stress Process	A N D	Measurement Detection Metrics Assessment Analysis Tracking	A N D	Measurement design strategy/methodology
	Goal: To determine the nature of auditory demands.						
2	Workload Work Quality Work Demands Work Process Effort	A N D	Mental Cognitive Stress	A N D	Measurement Detection Metrics Assessment Analysis Tracking Tools	A N D	Measurement design strategy/methodology
	Goal: To determine the presence of any and all measures of workload.						
3	Workload Work Quality Work Demands Work Process Effort	A N D	Method Methodology Criteria Assessment Analysis	A N D	Program(s) Tools Measurements Metrics	A N D	Measurement design strategy/methodology
	Goal: To determine the steps/process necessary to develop a workload assessment tool.						
4	1 embedded on/in 2	A N D		A N D	Auditory Workload Measures/Metrics Sensory Perception	A N D	Measurement design strategy/methodology
	Goal: To locate or determine the existence of any and all measures of auditory workload. This is the first focus of our *R&A*.						

ARL AUDITORY WORKLOAD *R&A* SEARCH STRATEGY							
	Main Ideas		Secondary Terms		R&A/Search Focus		Additional Items of Interest
5	2 embedded on/in 3	AND		AND	Processes Stages Program(s)	AND	Measurement design strategy/methodology
	Goal: To identify the steps/methodology necessary to establishing auditory workload measurements. This is the second focus of our *R&A*						
6	Sensory Perception Perceptual Sens* Percept*	AND	Processing Work Effort Demands Detection Coding Stress Process	AND	Tools Measurements Metrics Stages Program(s) Detection Assessment Analysis Tracking	AND	Measurement design strategy/methodology
	Goal: To further identify and tease out any additional methods of measuring sensory workload that may be of use in writing our *R&A*						

8. APPENDIX C: EXAMPLE OF STANDARD NASA-TLX INSTRUCTIONS

SUBJECT INSTRUCTIONS: RATINGS (Keyboard Version)

We are not only interested in assessing your performance but also the experiences you had during the different task conditions. Right now we are going to describe the technique that will be used to examine your experiences. In the most general sense we are examining the "Workload" you experienced. Workload is a difficult concept to define precisely, but a simple one to understand generally. The factors that influence your experience of workload may come from the task itself, your feelings about your own performance, how much effort you put in, or the stress and frustration you felt, The workload contributed by different task elements may change as you get more familiar with a task, perform easier or harder versions of it, or move from one task to another. Physical components of workload are relatively easy to conceptualize and evaluate. However, the mental components of workload may be more difficult to measure.

Since workload is something that is experienced individually by each person, there are no effective "rulers" that can be used to estimate the workload of different activities. One way to find out about workload is to ask people to describe the feelings they experienced. Because workload may be caused by many different factors, we would like you to evaluate several of them individually rather than lumping them into a single global evaluation of overall workload. This set of six rating scales was developed for you to use in evaluating your experiences during different tasks. Please read the descriptions of the scales carefully. If you have a question about any of the scales in the table, please ask me about it. It is extremely important that they be clear to you. You may keep the descriptions with you for reference during the experiment.

After performing each task, six rating scales will be displayed. You will evaluate the task by marking each scale at the point which matches your experience. Each line has two endpoint descriptors that describe the scale. Note that "own performance" goes from "good" on the left to "bad" on the right. This order has been confusing for some people. Move the arrow with the right and left arrow keys until it points at the desired location. Stop it by pressing the up arrow key. Press the down arrow key to enter your selection. Please consider your responses carefully in distinguishing among the task conditions. Consider each scale individually. Your ratings will play an important role in the evaluation being conducted, thus, your active participation is essential to the success of this experiment, and is greatly appreciated (NASA-TLX, v. 1.0).

About Human Systems IAC

The Human Systems Information Analysis Center (Human Systems IAC, HSIAC) is the gateway to worldwide sources of up-to-date human factors and ergonomics information and technologies for designers, engineers, researchers, and human factors specialists. Human Systems IAC provides a variety of products and services to government, industry, and academia while promoting the use of human factors and ergonomics in the design of human-operated equipment and systems.

Human Systems IAC's primary objective is to acquire, analyze, and disseminate timely information on human factors and ergonomics. In addition to providing free basic searches, Human Systems IAC performs other services on a cost-recovery basis:

- Distribute human factors and ergonomics technologies and publications
- Perform customized bibliographic searches and literature reviews
- Prepare state-of-the-art reports and critical review
- Conduct specialized analyses and evaluations
- Organize and conduct workshops and conferences

Human Systems IAC is a Department of Defense Information Analysis Center sponsored by the Defense Technical Information Center. It is technically managed by the Air Force Research Laboratory Human Effectiveness Directorate and operated by Booz Allen & Hamilton Inc.